

AD-A229 868

USAFSAM-TP-90-19

# MORBIDITY/MORTALITY ANALYSIS OF COHORT DATA

Daniel Mihalko, Ph.D. (Western Michigan University)

Southeastern Center for Electrical  
Engineering Education  
11th and Massachusetts Avenues  
St. Cloud, FL 34769

October 1990

DTIC  
ELECTE  
DEC 06 1990  
S E D

Final Report for Period March 1989 - March 1990

Approved for public release; distribution is unlimited.

Prepared for  
USAF SCHOOL OF AEROSPACE MEDICINE  
Human Systems Division (AFSC)  
Brooks Air Force Base, TX 78235-5301



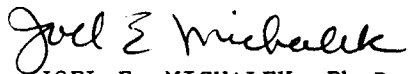
## NOTICES

This final report was submitted by the Southeastern Center for Electrical Engineering Education, 11th and Massachusetts Avenues, St. Cloud, Florida, under contract F33615-87-D-0609, job order SUPTXXEK, with the USAF School of Aerospace Medicine, Human Systems Division, AFSC, Brooks Air Force Base, Texas. Dr. Joel E. Michalek (SAM/EKB) was the Laboratory Project Scientist-in-Charge.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

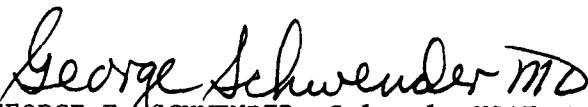
This report has been reviewed and is approved for publication.



JOEL E. MICHALEK, Ph.D.  
Project Scientist



WILLIAM H. WOLFE, Colonel, USAF, MC  
Chief, Epidemiology Division



GEORGE E. SCHWENDER, Colonel, USAF, MC, CFS  
Commander

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188		
1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS			
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.			
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE						
4. PERFORMING ORGANIZATION REPORT NUMBER(S)			5. MONITORING ORGANIZATION REPORT NUMBER(S)  USAFSAM-TP-90-19			
6a. NAME OF PERFORMING ORGANIZATION Southeastern Center for Elec- trical Engineering Education		6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION			
6c. ADDRESS (City, State, and ZIP Code) 11th and Massachusetts Avenues St. Cloud, FL 34769			7b. ADDRESS (City, State, and ZIP Code)			
8a. NAME OF FUNDING / SPONSORING ORGANIZATION USAF School of Aerospace Medicine		8b. OFFICE SYMBOL (If applicable) USAFSAM/EKB	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER  F33615-87-D-0609			
8c. ADDRESS (City, State, and ZIP Code) Human Systems Division (AFSC) Brooks Air Force Base, TX 78235-5301			10. SOURCE OF FUNDING NUMBERS			
			PROGRAM ELEMENT NO 87714F	PROJECT NO. SUPT	TASK NO XX	WORK UNIT ACCESSION NO. EK
11. TITLE (Include Security Classification)  Morbidity/Mortality Analysis of Cohort Data						
12. PERSONAL AUTHOR(S) Mihalko, Daniel (Western Michigan University, Kalamazoo MI 49008)						
13a. TYPE OF REPORT Final		13b. TIME COVERED FROM 89/3 TO 90/3	14. DATE OF REPORT (Year, Month, Day) 1990, October		15. PAGE COUNT 23	
16. SUPPLEMENTARY NOTATION						
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)			
FIELD	GROUP	SUB-GROUP	Maximum likelihood estimation, life expectancy Partial likelihood Survival analysis. (SS)			
12	03					
19. ABSTRACT (Continue on reverse if necessary and identify by block number)  This report presents three models for analyzing morbidity/mortality data. The three models are a fully nonparametric model, a parametric model, and a semiparametric model. Maximum likelihood methods are used to estimate the parameters in all three models. In addition, there is a partial likelihood solution for the semiparametric model. —						
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input type="checkbox"/> UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION Unclassified			
22a. NAME OF RESPONSIBLE INDIVIDUAL Joel E. Michalek, PhD			22b. TELEPHONE (Include Area Code) (512) 536-3441		22c. OFFICE SYMBOL USAFSAM/EKB	

## MORBIDITY/MORTALITY ANALYSIS OF COHORT DATA

### 1. INTRODUCTION

This report is the culmination of investigations into modeling the natural history of disease with application to the Air Force Health Study. Three approaches to the problem are presented here. The first approach extends the totally nonparametric method developed in Albert, Gertman and Louis (1978), Albert, Gertman, Louis, and Liu (1978) and Louis, Arthur, and Heghinian (1978). These three papers will be referred to as Louis et al. The second modeling approach is based on the semi-parametric methods for analyzing survival/sacrifice experiments presented in Dinse (1982, 1988) and Portier and Dinse (1987). This set of papers will be referred to as Dinse et al. The third approach combines the first two modeling methods to obtain a semi-parametric procedure using the disease model of Louis et al. and allows the inclusion of covariate data.

In Section 2 we present the extension of Louis et al. to include death as a third time-of-occurrence. In Section 3 we simplify the model and obtain a full likelihood solution. In Section 4 we show the drawbacks of attempting to extend the nonparametric model of Dinse (1982). In Section 5 we extend Dinse et al. to arrive at a parametric model. In Section 6 we present a partial likelihood solution to the problem of determining the effect of individual risk factors on death with disease and death without disease. In Section 7 we present the full likelihood solution. In a subsection of Section 7 we also present the formulas for a score test to determine the significance of risk factors.

### 2. LOUIS ET AL. EXTENSION TO MULTIPLE EXAMS

For multiple exams, the information available can be written in two parts: a basic Mortality/Morbidity vector denoted by MM and a matrix of examination data denoted by E. Let T = the age at death of a subject who has died, X = age at which the preclinical stage of the disease begins and Z = X + Y be the age at which symptoms appear. Y is the sojourn time in the preclinical stage.

The array MM is denoted by:

$$MM = (c, DELD, DS, S, C, DR, DC, I),$$

where

- c = age of subject at entrance into the study
- DS = min(T, time of analysis),
- DELD = 0 if DS = T,  
= 1 otherwise,
- Z = time of symptoms (of the disease of interest),
- S = min(time to symptoms, time of analysis, death without symptoms),
- C = 1 if subject is lost to follow-up at time S,  
= 0 if subject has shown symptoms at time S,
- DR = 1 if DS > X and DELD = 0,  
= 0 if DS ≤ X and DELD = 0,
- DC = 1 if DS ≥ X + Y and DELD = 0,  
= 0 if DS < X + Y and DELD = 0,

and I is the number of examinations (or screens) the subject received. DR and DC are undefined when DELD = 1. The array E is given by,  $E = (E(1), \dots, E(I))$ , where  $E(i) = (U(i), R(i))'$ , U(i) is the subject's age at the ith exam and R(i) = 0 if no disease is present at the ith exam and 1 otherwise.

The basic goal is to estimate the joint probability density function of  $(X, Z, T)$  in terms of MM and E. To illustrate the approach, we present here the calculation of the likelihood contribution of a particular realization of MM and E. Partition (or stratify) the positive real line into intervals  $I(1), \dots, I(M)$  and suppose that a subject's MM vector and E matrix are:

$$MM = (e \in I(a), DELD = 0, DS \in I(\ell), S \in I(k), C = 1, DR = 1, DC = 0, I = 2)$$

and

$$E = \begin{bmatrix} U(1) \in I(j_1), & R(1) = 0 \\ U(2) \in I(j_2), & R(2) = 1 \end{bmatrix}$$

Shorthand notation would be

$$MM = (a, 0, \ell, k, 1, 1, 0, 2)$$

and

$$E = \begin{bmatrix} j_1 & 0 \\ j_2 & 1 \end{bmatrix}$$

Let us review what information this data represents. In Figure 1 we abuse notation and treat an interval as a point in time. Figure 1 shows the disease and death history of a subject with this MM and E data.

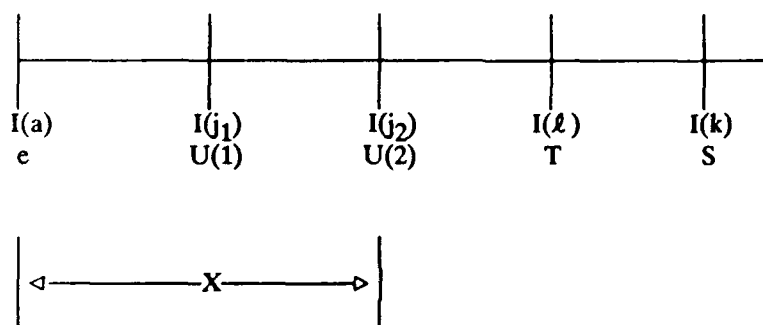


Figure 1.

Graphical representation of information available on a subject with age at entry ( $e$ ) in the time interval  $I(a)$ , age at death ( $T$ ) in the interval  $I(\ell)$ , age at first examination  $U(1)$  in the interval  $I(j_1)$ , age at second examination  $U(2)$  in the interval  $I(j_2)$ , and age at loss to follow-up ( $S$ ) in the interval  $I(k)$ .

In this special case, we know  $Z > T$ , since  $C = 1$ . We also know  $X < U(2)$ , because  $R(2) = 1$  (and we assume there are no false positives); however, if the screen at  $U(1)$  was a false negative,  $X$  may be less than  $U(1)$ . We also know  $DS = T$ , the actual age at death, since  $DELD = 0$ . Finally,  $I = 2$  tells us that this patient received two exams. Calculations for the likelihood contribution of this data are (with  $\rho$  = probability of a false negative result):



Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

$$\begin{aligned}
 P(MM,E) &= \sum_{q=1}^{j_2} P(X \in I(q), MM, E) \\
 &= \sum_{r=\ell+1}^M \sum_{q=1}^{j_2} P(X \in I(q), Z \in I(r), MM, E) \\
 &= \sum_{r=\ell+1}^M \sum_{q=1}^{j_2} P(\text{pos.screen at } I(j_2) \mid e \in I(a), X \in I(q), Z \in I(r), T \in I(\ell), E(1), I=2) \\
 &\quad \times P(e \in I(a), X \in I(q), Z \in I(r), T \in I(\ell), E(1))) \\
 &= \sum_{r=\ell+1}^M \sum_{q=1}^{j_2} 1 \times P(\text{neg.screen at } I(j_1) \mid e \in I(a), X \in I(q), Z \in I(r), T \in I(\ell), I=2) \\
 &\quad \times P(e \in I(a), X \in I(q), Z \in I(r), T \in I(\ell), I=2) \\
 &\quad \times P(U(1) \in I(j_1), U(2) \in I(j_2) \mid e \in I(a)) \\
 &\propto \sum_{r=\ell+1}^M \sum_{q=1}^{j_2} \rho^{\omega(j_1-q)} P(e \in I(a), X \in I(q), Z \in I(r), T \in I(\ell), I=2),
 \end{aligned}$$

where  $\omega(t) = 1 \ t > 0, \omega(t) = 0 \ t \leq 0$  and  $\propto$  indicates "proportional to".

Note that  $T \in I(\ell)$  and  $Z \in I(r)$ , with  $r \in \{\ell+1, \dots, M\}$ , is the information given by  $\{(C=1, S \in I(k), T \in I(\ell), \text{DELD}=0)\}$ . Under the assumption that examination times are scheduled regularly and independently of  $(X, Z, T)$ , as is the case in the Air Force Health Study, we have:

$$\begin{aligned}
 P(MM,E) &\propto \sum_{r=\ell+1}^M \sum_{q=1}^{j_2} \rho^{\omega(j_1-q)} P(a, q, r, \ell) P(I=2) \\
 &\propto \sum_{r=\ell+1}^M \sum_{q=1}^{j_2} \rho^{\omega(j_1-q)} P(a, q, r, \ell),
 \end{aligned}$$

where  $P(a, q, r, \ell) = P(e \in I(a), X \in I(q), Z \in I(r), T \in I(\ell))$ . The parameters  $p(a, q, r, \ell)$  are the same as those of the one-examination model. Thus, the Louis et al. model extended to include mortality data employs the same parameters in the case of two examinations as in the case of one examination.

### 3. SIMPLIFIED NONPARAMETRIC MODEL

In this simplified extension of Louis et al., we reduce the model by eliminating  $X$ . We demonstrate the technique by writing out the necessary equations in the special case that two exams have been administered and we use the language of the Air Force Health Study, since this form of the model is

specifically designed for the data from this study. Each subject's time is measured from his first tour in Vietnam. This simplified model uses only the following:

$T$  = time of death,  
 $Z$  = time of symptoms (of the disease of interest),  
 $DS = \min(T, \text{time of analysis}),$   
 $DELD = 0$  if  $DS = T,$   
 $= 1$  otherwise,  
 $S = \min(\text{time to symptoms, time to analysis or death without symptoms}),$   
 $DELZ = 1$  if  $S = Z,$   
 $= 0$  otherwise,  
 $I$  = number of exams,  
 $U(j)$  = time of the  $j$ th exam,  
 $R(j) = 1$  if  $U(j) \geq S,$   
 $= 0$  otherwise.

The time interval  $(0, \infty)$  is partitioned into intervals  $I(1), I(2), \dots, I(M)$ . Table 1 lists the possible data patterns (indexed by  $\eta$ ).

TABLE 1. POSSIBLE DATA PATTERNS IN THE REDUCED LOUIS ET AL. MODEL\*

Pattern No.	<u>S</u>	<u>DELZ</u>	<u>DS</u>	<u>DELD</u>	<u>I</u>	<u>U(1)</u>	<u>R(1)</u>	<u>U(2)</u>	<u>R(2)</u>
1	k	1	d	0	2	$j_1$	0	$j_2$	1
2	k	1	d	0	2	$j_1$	1	$j_2$	1
3	k	0	d	0	2	$j_1$	0	$j_2$	0
4	k	0	d	0	1	$j_1$	0	-	-
5	k	1	d	0	1	$j_1$	1	-	-
6	-	-	d	0	0	-	-	-	-
7	k	1	d	1	2	$j_1$	0	$j_2$	1
8	k	1	d	1	2	$j_1$	1	$j_2$	1
9	k	1	d	1	2	$j_1$	0	$j_2$	0
10	-	1	d	1	2	$j_1$	0	$j_2$	0
11	k	0	d	1	2	$j_1$	0	$j_2$	0
12	k	1	d	1	1	$j_1$	1	-	-
13	k	1	d	1	1	$j_1$	0	-	-
14	-	1	d	1	1	$j_1$	0	-	-
15	k	0	d	1	1	$j_1$	0	-	-
16	k	0	d	1	0	-	-	-	-
17	k	1	d	1	0	-	-	-	-
18	-	1	d	1	0	-	-	-	-

\*Lower case letters (k, d,  $j_1, j_2 = 1, \dots, M$ ) in the Z, DS, U(1), U(2) columns indicate interval numbers in the partition of  $(0, \infty)$ .

The goal of this nonparametric approach is to estimate the joint distribution of  $(Z, T)$ . Table 2 lists the likelihood contribution of each data pattern in terms of  $P(a, b) = P(Z \in I(a), T \in I(b))$ .

TABLE 2. LIKELIHOOD CONTRIBUTION OF DATA PATTERNS IN THE REDUCED LOUIS ET AL. MODEL\*

Data Pattern $\eta$	Likelihood Contribution
1	$\sum_{\ell=d+1}^M P(k, \ell)$
2	$\sum_{\ell=d+1}^M P(k, \ell)$
3	$\sum_{q=j_2+1}^M \sum_{\ell=d+1}^M P(q, \ell)$
4	$\sum_{q=j_1+1}^M \sum_{\ell=d+1}^M P(q, \ell)$
5	$\sum_{\ell=d+1}^M P(k, \ell)$
6	$\sum_{q=1}^M \sum_{\ell=d+1}^M P(q, \ell)$
7	$P(k, d)$
8	$P(k, d)$
9	$P(k, d)$
10	$\sum_{q=j_2+1}^d P(q, d)$
11	$\sum_{q=j_2+1}^M P(q, d)$
12	$P(k, d)$
13	$P(k, d)$
14	$\sum_{q=j_1+1}^d P(q, d)$
15	$\sum_{q=d+1}^M P(q, d)$
16	$\sum_{q=d+1}^M P(q, d)$
17	$P(k, d)$
18	$\sum_{q=1}^d \sum_{\ell=d+1}^M P(q, \ell)$

\* $P(a,b) = P(Z \in I(a), T \in I(b))$

We wish to maximize the resulting likelihood in terms of the parameters  $P(a, b)$ . Because the likelihood function is too complicated to maximize directly, we apply (as did Louis et al.) the Estimation and Maximization (EM) algorithm to obtain maximum likelihood estimates of the  $P(a, b)$ . To invoke the EM algorithm we need to compute the conditional probabilities  $P(Z \in I(c), T \in I(f) | \eta)$ , the proba-



bility that  $Z \in I(c)$  and  $T \in I(f)$  given that the subject has displayed data-pattern  $\eta$ . Using the definition of conditional probability we see that

$$P(Z \in I(c), T \in I(f) | \eta) = P(Z \in I(c), T \in I(f), \eta) / P(\eta), \eta = 1, 2, \dots, 18,$$

where  $P(\eta)$  is the likelihood contribution of the data pattern  $\eta$ . To compute these conditional probabilities we first compute the probabilities of a subject having  $Z \in I(c)$  (ie, symptoms occur in the interval  $I(c)$ ) and  $T \in I(f)$  (death occurs in interval  $I(f)$ ) and the subject displays data pattern  $\eta$ . Let

$$\delta(t) = 1 \text{ if } t = 0 \\ = 0 \text{ otherwise,}$$

and

$$\omega(t) = 1 \text{ if } t > 0 \\ = 0 \text{ otherwise}$$

Table 3 gives these probabilities for each  $\eta, \eta = 1, 2, \dots, 18$ .

TABLE 3. THE PROBABILITIES THAT SUBJECT'S SYMPTOMS APPEAR AT TIME S IN THE INTERVAL  $I(C)$  AND THAT THE SUBJECT'S TIME OF DEATH (T) IS IN THE INTERVAL  $I(F)$ , GIVEN THAT THE SUBJECT HAS DATA PATTERN  $\eta$

$\eta$	$(S, \text{DELZ}, \text{DS}, \text{DELD}, I, U(1), R(1), U(2), R(2))$	$P(Z \in I(c), T \in I(f), \eta)$
1	$(k, 1, d, 0, 2, j_1, 1, j_2, 0)$	$P(c, f) \omega(f-d) \delta(k-c)$
2	$(k, 1, d, 0, 2, j_1, 1, j_2, 1)$	$P(c, f) \omega(f-d) \delta(k-c)$
3	$(k, 0, d, 0, 2, j_1, 0, j_2, 0)$	$P(c, f) \omega(c-j_2) \omega(f-d)$
4	$(k, 0, d, 0, 1, j_1, 0, -, -)$	$P(c, f) \omega(c-j_1) \omega(f-d)$
5	$(k, 1, d, 0, 1, j_1, 1, -, -)$	$P(c, f) \omega(f-d) \delta(k-c)$
6	$(-, -, d, 0, 0, -, -, -, -)$	$P(c, f) \omega(f-d)$
7	$(k, 1, d, 1, 2, j_1, 0, j_2, 1)$	$P(c, f) \delta(f-d) \delta(k-c)$
8	$(k, 1, d, 1, 2, j_1, 1, j_2, 1)$	$P(c, f) \delta(f-d) \delta(k-c)$
9	$(k, 1, d, 1, 2, j_1, 0, j_2, 0)$	$P(c, f) \delta(f-d) \delta(k-c)$
10	$(-, 1, d, 1, 2, j_1, 0, j_2, 0)$	$P(c, f) \omega(c-j_2) \delta(f-d) (1-\omega(d-c))$
11	$(j_2, 0, d, 1, 2, j_1, 0, j_2, 0)$	$P(c, f) \omega(c-d) \delta(f-d)$
12	$(k, 1, d, 1, 1, j_1, 1, -, -)$	$P(c, f) \delta(k-c) \delta(f-d)$
13	$(k, 1, d, 1, 1, j_1, 0, -, -)$	$P(c, f) \delta(k-c) \delta(f-d)$
14	$(-, 1, d, 1, 1, j_1, 0, -, -)$	$P(c, f) \omega(c-j_1) \delta(f-d) (1-\omega(d-c))$
15	$(d, 0, d, 1, 1, j_1, 0, -, -)$	$P(c, f) \omega(c-d) \delta(f-d)$
16	$(k, 0, d, 1, 0, -, -, -, -)$	$P(c, f) \omega(c-k) \delta(f-d)$
17	$(k, 1, d, 1, 0, -, -, -, -)$	$P(c, f) \delta(k-c) \delta(f-d)$
18	$(k, 1, d, 1, 0, -, -, -, -)$	$P(c, f) (1-\omega(d-c)) \delta(f-d)$

The maximum likelihood estimates of the  $P(a, b)$  are obtained by using the following EM algorithm. Let  $n(\eta)$  = the number of subjects with data pattern  $\eta, \eta = 1, 2, \dots, 18$ .

Step 1: Choose a set of initial probabilities:

$$P^{(0)}(a, b), a, b = 1, \dots, M.$$

Step 2: Given the  $\nu$ th estimates  $P^{(\nu)}(a, b), \nu = 0, 1, 2, \dots$ , compute

$$n^{(\nu+1)}(a,b) = \sum_{\eta=1}^{18} n(\eta) P^{(\nu)}(a,b | \eta), a,b=1,\dots,M,$$

where  $P^{(\nu)}(a,b | \eta)$  is obtained by using  $P^{(\nu)}(a,b)$  in the formulas of Tables 1 and 3 and the formula for  $P(Z \in I(a), T \in I(b) | \eta)$ .

Step 3: Compute  $P^{(\nu+1)}(a,b) = n^{(\nu+1)}(a,b)/N$ ,  $a,b=1,\dots,M$ , where  $N = \sum_{\eta} n(\eta)$ .

Step 4: Repeat Steps 2 and 3 until the  $P^{(\nu)}(a,b)$  converge.

The resulting limits are the maximum likelihood estimates of the  $P(a,b)$ .

#### 4. TWO-EXAMINATION EXTENSION OF DINSE (1982) AND ITS DRAWBACKS

Dinse (1982) provides a method for analyzing data composed of examination and death time information when subjects are examined at most one time. Below we provide an example of an extension of this work that allows more than one examination. This development is exactly that of Dinse (1982) except that the number of examinations is 2 rather than 1. This approach illustrates the fact that, unlike the extension of Louis et al. in Section 2, increasing the number of examinations by one increases the number of parameters in the model by a number equal to the product of the number of death times and the number of disease states. This increase in parameters makes the method too cumbersome for a study like the Air Force Health Study in which there are to be up to six examinations. The full data case is sufficient to demonstrate the drawback of attempting to use a two-examination extension of Dinse.

Define the following notation:

$T$  = age at death,  
 $B(T)$  = disease state at death,  
 $E(i)$  = age at the  $i$ th exam,  $i = 1, 2$ ,  
 $B(i)$  = disease state at the  $i$ th exam.

The full data case involves no incomplete pairs among  $(T, B(T))$ ,  $(E(i), B(i))$ , although one or more of  $(E(i), B(i))$  may be missing. By design  $E(1) < E(2)$ , so that there are three possibilities for an individual:  $T < E(1) < E(2)$ ,  $E(1) < T < E(2)$ , or  $E(1) < E(2) < T$ . If  $T < E(i)$ , then the  $i$ th examination was not performed (of course).

If we follow the nonparametric approach of Dinse, we would take each possible event and write out its likelihood contribution as a product of conditional and unconditional probabilities. These conditional and unconditional probabilities would then be considered parameters to be estimated. The infeasibility of this approach for the Air Force Health Study can be best illustrated by writing out the likelihood contribution for an individual having had two exams (ie,  $E(1) < E(2) < T$ ). Using the time intervals  $I(1) < I(2) < \dots < I(M)$  of the Louis et al. extension, the likelihood contribution would be:

$$\begin{aligned} &P\{T \in I(a), B(T) = b(T), E(1) \in I(c), B(1) = b(1), E(2) \in I(d), B(2) = b(2)\} \\ &= P(B(T) = b(T) | T \in I(a), E(1) \in I(c), B(1) = b(1), E(2) \in I(d), B(2) = b(2)) \\ &\times P(T \in I(a) | E(1) \in I(c), B(1) = b(1), E(2) \in I(d), B(2) = b(2)) \times P(B(2) = b(2) | E(1) \in I(c), \\ &\quad B(1) = b(1), E(2) \in I(d)) \times P(B(1) = b(1) | E(1) \in I(c)) P(E(1) \in I(c), E(2) \in I(d)). \end{aligned}$$

The first factor requires a parameter for every combination of death time, exam times and disease states. Essentially, this parameter forces the data to be stratified according to each such combination.

This problem illuminates the difference between survival/sacrifice experiments and epidemiological studies. In a survival/sacrifice experiment the sacrifice is the first and only examination and coincides with the ending of information on that animal. However, in an epidemiologic study with repeated exams, the subjects may continue to live after the first and second exams and their disease history will take one of many paths. To make inferences about these paths, we must have a sufficient number of both cases and controls taking each path. A large number of paths, such as would be the situation for two or more exams, would require a very large and therefore infeasible number of cases and controls. Clearly, this is an unrealistic approach for the Air Force Health Study. An alternative and feasible approach is to invoke a parametric model, described in the next section.

## 5. A PARAMETRIC MODEL

Now let  $X$  denote the time to the first event, either onset of the disease of interest or death without the disease of interest. Let  $T$  denote the time to natural death. Define

$$Y(t) = 1 \text{ if disease is present at time } t, \\ = 0 \text{ otherwise.}$$

Suppose there are  $J$  distinct death and examination times. Let  $V$  be a vector of covariates measured on each subject. Define

$a_j(V)$  = number of subjects who died with the disease at  $t_j$  and had covariate vector  $V$ .

$b_j(V)$  = number of subjects who died without the disease at  $t_j$  and had covariate vector  $V$ .

$m_j(V)$  = number of subjects examined at  $t_j$  who were disease free and had covariate vector  $V$ .

$n_j(V)$  = number of subjects examined at  $t_j$  who had the disease and had covariate vector  $V$ .

Define the following hazard functions:

$$\lambda_v(t) = \lim_{\epsilon \rightarrow 0} P(t \leq X < t + \epsilon, Y(X) = 1 \mid X \geq t, V = v) / \epsilon,$$

$$\beta_v(t) = \lim_{\epsilon \rightarrow 0} P(t \leq X < t + \epsilon, Y(X) = 0 \mid X \geq t, V = v) / \epsilon,$$

$$= \lim_{\epsilon \rightarrow 0} P(t \leq T < t + \epsilon \mid T \geq t, Y(t) = 0, V = v) / \epsilon,$$

$$\alpha_v(t) = \lim_{\epsilon \rightarrow 0} P(t \leq T < t + \epsilon \mid T \geq t, Y(t) = 1, V = v) / \epsilon.$$

These functions extend Dinse's (1988) expressions to include covariates. It follows directly from Dinse (1988) that  $\lambda_v(t)$ ,  $\beta_v(t)$  and  $\alpha_v(t)$  can all be written in terms of the following three functions:

$$h_v(t) = \lim_{\epsilon \rightarrow 0} P(t \leq T < t + \epsilon \mid T \geq t, V = v) / \epsilon,$$

$$\pi_v(t) = P(Y(t) = 1 \mid T \geq t, V = v),$$

and

$$p_v(t) = P(Y(t) = 1 \mid T = t, V = v).$$

Following Dinse (1988), we define the lethality function as  $r_v(t) = \alpha_v(t) / \beta_v(t)$ . It follows from Dinse's (1988) expression (4) that  $r_v(t) = (p_v(t) / (1 - p_v(t))) / (\pi_v(t) / (1 - \pi_v(t)))$ .

We now introduce the following model: Let  $\xi$ ,  $\nu$  and  $\gamma$  be vectors of coefficients and  $\xi_0(t)$ ,  $\nu_0(t)$  and  $\gamma_0(t)$  be functions of time. We model  $h_1(t) = h_0(t)e^{\xi'v}$ , a proportional hazards model, and  $\pi_v(t)$  and  $p_v(t)$  with logistic models,

$$\pi_v(t) = \frac{e^{\nu_0(t) + \nu'v}}{1 + e^{\nu_0(t) + \nu'v}}$$

and

$$p_v(t) = \frac{e^{\gamma_0(t) + \gamma'v}}{1 + e^{\gamma_0(t) + \gamma'v}}.$$

It follows that the survival function  $S_v(t) = P_v(T > t)$  can be written as a power of a baseline survival function determined by  $h_0(t)$ ,

$$\begin{aligned} S_v(t) &= \exp\left\{-\int_0^t h_v(u) du\right\} \\ &= \left(\exp\left\{-\int_0^t h_0(u) du\right\}\right) e^{\xi'v}. \end{aligned}$$

We also see that

$$r_v(t) = e^{\gamma_0(t) - \nu_0(t) + (\gamma' - \nu')v},$$

so that the ratio of lethality for covariate vectors  $V = v_1$  to  $V = v_2$  is

$$r_{v_1}(t) / r_{v_2}(t) = e^{(\gamma' - \nu')(v_1 - v_2)}.$$

The theory in section 3 of Dinse (1988) together with the above models imply that the loglikelihood is the sum over all vectors  $z$  of the following expressions.

$$L_1(h_v) = \sum_{j=1}^J \{(a_j(v) + b_j(v))(\log h_0(t_j) + \xi'v) - (a_j(v) + b_j(v) + m_j(v) + n_j(v))e^{\xi'v} \int_0^{t_j} h_0(u) du\}$$

$$L_2(\pi_v) = \sum_{j=1}^J \{n_j(v)(\nu_0(t_j) + \nu'v) - (n_j(v) + m_j(v))\log(1 + e^{\nu_0(t_j) + \nu'v})\}$$

$$L_3(p_v) = \sum_{j=1}^J \{b_j(v)(\gamma_0(t_j) + \gamma'v) - (a_j(v) + b_j(v))\log(1 + e^{\gamma_0(t_j) + \gamma'v})\}.$$

The resulting loglikelihood is  $L = \sum_v [L_1(h_v) + L_2(\pi_v) + L_3(p_v)]$ .

The baseline information  $h_0(t)$ ,  $\nu_0(t)$  and  $\gamma_0(t)$  will be known if the exposed group is being compared with a population (Breslow, Lubin, Marek, and Langholz (1983)). In a two-sample study, however, this baseline information will not be known and thus must be estimated. Such estimates are derived via modeling. Survival times are frequently assumed to follow a Weibull model (Dinse 1988). Dinse and Lagakos (1982) suggest modeling the logit of the prevalence rate by a low-order polynomial in time. Taking these suggestions, we model

$$h_0(t) = \mu_0 \mu_1 t^{\mu_1 - 1}, \quad (1)$$

$$\nu_0(t) = \nu_0 + \nu_1 t + \nu_2 t^2$$

and

$$\gamma_0(t) = g_0 + g_1 t + g_2 t^2.$$

The resulting parameters are estimated by maximizing the full likelihood.

### 5.1 Some Special Cases

In the following special cases we assume  $v$  is a group indicator,  $v = 1$  for exposed and 0 for controls.

Case 1:  $p_v(t) = p_0(t)$ ,  $\pi_v(t) = \pi_0(t)$  and  $h_0(t)$  is known.

This is the case assumed in Breslow, Lubin, Marek, and Langholz (1983). In this situation the only difference between exposed and controls is in survival. The loglikelihood in this situation is  $L = \sum_v L_1(h_v)$ . The hypothesis of interest is  $H_0: \xi = 0$ . The corresponding score test (Rao, 1973) which is based on the statistic  $S = (\partial L / \partial \xi | \xi = 0)^2 / (-\partial^2 L / \partial \xi^2 | \xi = 0)$ . In this case,

$$\begin{aligned} \partial L / \partial \xi | \xi = 0 &= \sum_{j=1}^J \{(a_j(1) + b_j(1)) - (a_j(1) + b_j(1) + n_j(1)) \int_0^{t_j} h_0(u) du\}, \\ &= O-E \end{aligned}$$

where  $O = \sum_{j=1}^J (a_j(1) + b_j(1)) =$  number of deaths among the exposed group and

$$E = \sum_{j=1}^J \{(a_j(1) + b_j(1) + m_j(1) + n_j(1)) \int_0^{t_j} h_0(u) du\}$$

= expected number of deaths among the exposed group  
under the control hazard,

and  $(-\partial^2 L / \partial \xi^2) |_{\xi=0} = E$ . Rao's score statistic for testing  $H_0: \xi = 0$  is thus  $(O-E)^2/E$ . This is the same test statistic obtained by Breslow, Lubin, Marek, and Langholz (1983) in a similar situation.  
Case 2:  $\xi = 0, \gamma = 0$  with  $\nu_0(t)$  and  $\gamma_0(t)$  known.

This is the case in which exposed and controls differ only in the prevalence of disease among the living, with the prevalence among the living and prevalence at death assumed known for the controls. In this case the observed number of diseased subjects among the living exposed is:

$$O = \sum_{j=1}^J n_j(1),$$

while the expected number of disease subjects among the exposed under the control prevalence rate is:

$$E = \sum_{j=1}^J (m_j(1) + n_j(1)) (e^{\nu_0(t_j) + \nu} / (1 + e^{\nu_0(t_j) + \nu})).$$

Hence,  $(\partial L / \partial \nu |_{\nu=0}) = O-E$  and  $(\partial^2 L / \partial \nu^2 |_{\nu=0}) = -E$ . Thus the score statistic for testing  $H_0: \nu = 0$  is again of the form  $(O-E)^2/E$ .

Case 3:  $\xi = 0, \nu = 0$  with  $\nu_0(t)$  and  $\gamma_0(t)$  known.

In this case the only difference between exposed and controls is the prevalence of disease at death, with the control prevalence known. It again follows that the score statistic for testing  $H_0: \gamma = 0$  is  $(O-E)^2/E$ , where

$$O = \sum_{j=1}^J b_j(1) = \text{observed number of diseased cases among those dying of natural causes,}$$

and

$$E = \sum_{j=1}^J (a_j(1) + b_j(1)) (e^{\gamma_0(t_j) + \gamma} / (1 + e^{\gamma_0(t_j) + \gamma})).$$

## 5.2 The General Case

In the general case that the baseline functions cannot be assumed known and there may be covariate effects for survival and both prevalences, we assume the baseline models in (1). The loglikeli-

hood is the sum over  $z$  of  $L_1(h_z)$ ,  $L_2(\pi_v)$  and  $L_3(p_v)$  with (1) substituted for  $h_0(t)$  and  $\nu_0(t)$  and  $\gamma_0(t)$ . The resulting likelihood is then maximized to obtain the parameter estimates.

## 6. A PARTIAL LIKELIHOOD MODEL

Again, let  $Z$  indicate the time symptoms occurred and let  $T$  indicate the time of death. In this section, we extend the Dinse (1982) model in terms of  $Z$  and  $T$  and then use this model to build a likelihood. The events for which we need likelihood contributions, numbered I, II, III, IV, and V as in Dinse (1982), are all the combinations of censoring situations on  $Z$  and  $T$ . Table 4 lists these events and their likelihood contributions.

TABLE 4. OBSERVABLE EVENTS AND THEIR LIKELIHOOD CONTRIBUTIONS FOR THE DINSE (1982) EXTENSION

Event Type	Event	Likelihood Contribu
I	$Z > t, T > t$	$P(T > t   Z > t)P(Z > t)$
II	$Z < t, T > t$	$P(T > t   Z < t)P(Z < t)$
III	$Z > t, T = t$	$-\frac{d}{dt} P(T > t   Z > t)P(Z > t)$
IV	$Z < t, T = t$	$-\frac{d}{dt} P(T > t   Z < t)P(Z < t)$
V	$T > t$	Sum of likelihood contributions I, II

Notice that we are assuming that the actual time of occurrence of symptoms is never known. The model can be easily extended to include such cases if necessary. Notice also that the events depict knowledge at a point in time  $T = t$ . This time  $T$  would be the time of the last available knowledge about the subject. Thus  $T$  can be the time of death, the time of loss to follow-up, or time of analysis.

Let  $V$  indicate a vector of covariates. We define the following conditional hazards.

$$\lambda(t | Z < t, V = v) = \lim_{\epsilon \rightarrow 0} P(t \leq T < t + \epsilon | Z < t, V = v, T \geq t) / \epsilon$$

and

$$\lambda(t | Z > t, V = v) = \lim_{\epsilon \rightarrow 0} P(t \leq T < t + \epsilon | Z > t, V = v, T \geq t) / \epsilon.$$

We also define the conditional baseline hazards as:

and

$$\lambda_1(t) = \lambda(t | Z < t, V = 0)$$

$$\lambda_2(t) = \lambda(t | Z > t, V = 0).$$

Finally, we model the conditional hazards of T as:

$$\lambda(t | Z < t, V = v) = \lambda_1(t) e^{\beta' v}$$

and

$$\lambda(t | Z > t, V = v) = \lambda_2(t) e^{\alpha' v},$$

where  $\beta$  and  $\alpha$  are vectors of parameters which reflect the effect of the covariates in V on the hazard of death for a person with and without the disease, respectively.

Let  $t_{(1)} < t_{(2)} < \dots < t_{(k)}$  be the ordered death times of those dying with the disease. Denote by  $V_{(j)}$  the vector of covariates for the subject dying at  $t_{(j)}$ . Similarly, let  $\bar{t}_{(1)} < \bar{t}_{(2)} < \dots < \bar{t}_{(k)}$  be ordered death times of those dying without the disease. Let  $\bar{V}_{(j)}$  be the vector of covariates for the subject dying at  $\bar{t}_{(j)}$ . Define  $R_{(j_1)}$  and  $\bar{R}_{(j_2)}$  to be the set of indices for subjects at risk at  $t_{(j_1)}$  and  $\bar{t}_{(j_2)}$ , respectively,  $j_1 = 1, 2, \dots, k$ ,  $j_2 = 1, 2, \dots, k$ . Standard partial likelihood arguments (Lawless (1982), pp356-357) lead to the partial likelihood

$$L(\alpha, \beta) = \left( \prod_{j_1=1}^k \frac{e^{\beta' V_{(j_1)}}}{\sum_{\ell \in R_{(j_1)}} e^{\beta' V_{(\ell)}}} \right) \left( \prod_{j_2=1}^k \frac{e^{\alpha' \bar{V}_{(j_2)}}}{\sum_{\ell \in \bar{R}_{(j_2)}} e^{\alpha' \bar{V}_{(\ell)}}} \right). \quad (2)$$

This partial likelihood is the product of two standard partial likelihood functions. Thus, it can be maximized using standard methods on each factor to derive estimates of  $\alpha$  and  $\beta$ .

## 7. A FULL LIKELIHOOD MODEL

The partial likelihood solution to (2) allows us to compare the effect on death rates of a covariate for subjects with and without the disease. However, there are other descriptors for the disease which we cannot estimate without further modeling. For example, the lethality of the disease at age t for subjects with covariates  $V = v$  can be defined as:

$$r_v = \frac{\lambda(t | S < t, V = v)}{\lambda(t | S > t, V = v)}.$$

In terms of the model in Section 2 we have:

$$r_v = \frac{\lambda_1(t)}{\lambda_2(t)} e^{(\beta - \alpha)' v},$$

so that, without a trivializing assumption on the baseline hazards, we can make no inference about  $r_v(t)$ . The solution to this problem is to model the baseline hazards. In this section, we give a full likelihood solution to the discretized form of this problem. This solution will be useful in analyzing grouped data.



Let  $t_1, t_2, \dots, t_k$  be the set of possible death times. Define:

$$\lambda_{1j} = P(T = t_j \mid T \geq t_j, S \leq t_j), j = 1, \dots, k$$

and

$$\lambda_{2j} = P(T = t_j \mid T \geq t_j, S > t_j), j = 1, \dots, k.$$

Let

$D_j(d)$  = The set of labels of individuals dying at time  $t_j$  and with the disease ( $S \leq t_j$ )

$C_j(d)$  = The set of labels of those with the disease censored at  $t_j$ .

$D_j(\bar{d})$  = The set of labels of subjects without the disease who died at  $t_j$ .

$C_j(\bar{d})$  = The set of labels of subjects without the disease who were censored at  $t_j$ .

$C_j$  = The set of labels of persons whose death times were censored at  $t_j$  and for whom we have no disease information.

Following the technique of Kalbfleisch and Prentice (1980), pp 98-102, we model:

$$\gamma_{1j} = \log((- \log(1 - \lambda_{1j}))$$

and

$$\gamma_{2j} = \log((- \log(1 - \lambda_{2j}))). \quad (3)$$

Using the discrete form of the proportional hazards model, we have:

$$P(T = t_j, S \leq t_j \mid T \geq t_j, V = v) = 1 - (1 - \lambda_{1j}) e^{\beta'v}$$

$$P(T = t_j, S > t_j \mid T \geq t_j, V = v) = 1 - (1 - (1 - \lambda_{2j})) e^{\alpha'v}.$$

The resulting likelihood is:

$$\begin{aligned} L(\lambda, \alpha, \beta) = & \prod_{j=1}^k \left\{ \prod_{i \in D_j(d)} [1 - (1 - \lambda_{1j}) e^{\beta'v_i}] \prod_{\ell \in C_j(d)} (1 - \lambda_{1j}) e^{\beta'v_\ell} \prod_{i \in D_j(\bar{d})} [1 - (1 - \lambda_{2j}) e^{\alpha'v_i}] \right. \\ & \left. \times \prod_{\ell \in C_j(\bar{d})} (1 - \lambda_{2j}) e^{\alpha'v_\ell} \prod_{\ell \in C_j} [1 - (1 - \lambda_{1j}) e^{\beta'v_\ell} + (1 - \lambda_{2j}) e^{\alpha'v_\ell}] \right\}, \quad (4) \end{aligned}$$

where  $\lambda = (\lambda_{11}, \dots, \lambda_{1k}, \lambda_{21}, \dots, \lambda_{2k})'$ .

If there are no censored death times without disease information, then  $L(\lambda, \alpha, \beta)$  is the product of two standard likelihoods for the discrete proportional hazards model. Substituting (3) into (4) and taking the logarithm, allows us to write the log likelihood as:

$$\log L(\gamma, \alpha, \beta) = \sum_{j=1}^k \left\{ \sum_{i \in D_j(d)} \log(1 - \exp(-\exp(\gamma_{1j} + \beta'v_i))) - \sum_{\ell \in C_j(d)} \exp(\gamma_{1j} + \beta'v_\ell) \right\}$$

$$\begin{aligned}
& + \sum_{i \in D_j(d)} \log(1 - \exp(-\exp(\gamma_{2j} + \alpha' v_i))) - \sum_{\ell \in C_j(d)} \exp(\gamma_{2j} + \alpha' v_\ell) \\
& + \sum_{\ell \in C_j} \log(\exp(-\exp(\gamma_{1j} + \beta' v_\ell)) + \exp(-\exp(\gamma_{2j} + \alpha' v_\ell))).
\end{aligned}$$

To form the score test for  $H_0: \beta = 0, \alpha = 0$ , we compute for  $j = 1, \dots, k$ ,  $j_1 = 1, 2, \dots, k$ ,  $j_2 = 1, 2, \dots, k$ ,  $\nu = 1, 2$  and  $r, q = 1, \dots, p$ ,

$$[\partial \Gamma_j] = \left\{ \frac{\partial}{\partial \gamma_{ij}} \log L \right\},$$

$$[\partial \beta] = \left\{ \frac{\partial}{\partial \beta_r} \log L \right\},$$

$$[\partial \alpha] = \left\{ \frac{\partial}{\partial \alpha_r} \log L \right\},$$

$$[-\partial^2 \beta] = \left\{ \frac{-\partial^2 \log L}{\partial \beta_r \partial \beta_q} \right\},$$

$$[-\partial^2 \alpha] = \left\{ \frac{-\partial^2 \log L}{\partial \alpha_r \partial \alpha_q} \right\},$$

$$[-\partial^2 \beta \alpha] = \left\{ \frac{-\partial^2 \log L}{\partial \alpha_r \partial \beta_q} \right\},$$

$$[-\partial^2 \Gamma_j \beta] = \left\{ \frac{-\partial^2 \log L}{\partial \gamma_{ij} \partial \beta_r} \right\},$$

$$[-\partial^2 \Gamma_j \alpha] = \left\{ \frac{-\partial^2 \log L}{\partial \gamma_{ij} \partial \alpha_r} \right\},$$

and

$$[-\partial^2 \Gamma_{j_1} \Gamma_{j_2}] = \left\{ \frac{-\partial^2 \log L}{\partial \gamma_{1j_1} \partial \gamma_{2j_2}} \right\}.$$

$[\partial \Gamma_j]$  is a  $k \times 1$  vector with  $j$ th component  $\frac{\partial}{\partial \gamma_{ij}} \log L$ ,  $[-\partial^2 \beta \alpha]$  is a  $p \times p$  matrix and  $[-\partial^2 \Gamma_j \beta]$  is a  $k \times p$  matrix,  $i = 1, 2$ . Define  $\hat{\Gamma}$  to be the vector of  $\hat{\gamma}_{ij}$ , the solutions to  $\frac{\partial \log L}{\partial \gamma_{ij}} = 0$  where the derivative is evaluated at  $\alpha = 0$  and  $\beta = 0$ . The components of  $\hat{\Gamma}$  can be written simply with more notation.

Define:

$a_{1j}$  = The number of indices in  $D_j(d)$ ,

$a_{2j}$  = The number of indices in  $D_j(\bar{d})$ ,

$b_{1j}$  = The number of indices in  $C_j(d)$ ,

$b_{2j}$  = The number of indices in  $C_j(\bar{d})$ ,

and

$c_j$  = The number of indices in  $C_j$ .

Finally define:

$$p_{\nu j} = \exp[-\exp(\gamma_{\nu j})], \quad j = 1, \dots, k, \quad \nu = 1, 2.$$

With this notation the components of  $\hat{\Gamma}$  are the solutions to the following  $k$  pairs of quadratic equations.

$$p_{1j}^2(a_{1j} + b_{1j} + c_j) + p_{1j}(-b_{1j} - c_j) + p_{1j}p_{2j}(a_{1j} + b_{1j}) - p_{2j}b_{1j} = 0$$

and

$$p_{2j}^2(a_{2j} + b_{2j} + c_j) + p_{2j}(-b_{2j} - c_j) + p_{1j}p_{2j}(a_{2j} + b_{2j}) - p_{1j}b_{2j} = 0,$$

$j = 1, \dots, k$ . Define:

$$(\partial L)' = ([\partial \Gamma_1]', [\partial \Gamma_2]', [\partial \beta]', [\partial \alpha]')$$

and  $\Sigma$  as the  $2(k+p) \times 2(k+p)$  symmetric matrix

$$\begin{bmatrix} [-\partial^2 \Gamma_1] & [-\partial^2 \Gamma_1 \Gamma_2] & [-\partial^2 \Gamma_1 \beta] & [-\partial^2 \Gamma_1 \alpha] \\ & [-\partial^2 \Gamma_2] & [-\partial^2 \Gamma_2 \beta] & [-\partial^2 \Gamma_2 \alpha] \\ & & [-\partial^2 \beta] & [-\partial^2 \beta \alpha] \\ & & & [-\partial^2 \alpha] \end{bmatrix}$$

Then, the score test for  $H_0: \beta = 0, \alpha = 0$  is the  $X_{2p}^2$  statistic

$$X^2 = (\partial L)_0' \hat{\Sigma}_0^{-1} (\partial L)_0,$$

where the subscript "0" indicates evaluation at  $(\Gamma, \alpha, \beta) = (\hat{\Gamma}, 0, 0)$ .

### 7.1 Computational Formulas For The Score Test

The following formulas are all evaluated at  $\alpha = 0, \beta = 0$ . The  $r$ th component of the vector of covariates for the  $i$ th subject is denoted by  $Z_{ir}$ .

$$\frac{\partial \log L}{\partial \gamma_{\nu i}} = \frac{a_{\nu j} p_{\nu j}}{1 - p_{\nu j}} - b_{\nu j} - \frac{c_j p_{\nu j}}{p_{1j} + p_{2j}}, \quad j = 1, \dots, K, \quad \nu = 1, 2.$$

$$\frac{\partial \log L}{\partial \beta_r} = \sum_{j=1}^k \left\{ \frac{-p_{1j} \log p_{1j}}{1 - p_{1j}} \sum_{i \in C_j(d)} z_{ir} + \log p_{1j} \sum_{\ell \in C_j(d)} z_{\ell r} + \frac{p_{1j} \log p_{1j}}{p_{1j} + p_{2j}} \sum_{\ell \in C_j} z_{\ell r} \right\}, \quad r = 1, \dots, p.$$

$$\frac{\partial \log L}{\partial \alpha_r} = \sum_{j=1}^k \left\{ \frac{-p_{2j} \log p_{2j}}{1 - p_{2j}} \sum_{i \in D_j(d)} z_{ir} + \log p_{2j} \sum_{\ell \in C_j(d)} z_{\ell r} + \frac{p_{2j} \log p_{2j}}{p_{1j} + p_{2j}} \sum_{\ell \in C_j} z_{\ell r} \right\}, \quad r = 1, \dots, p.$$

$$\begin{aligned} \frac{-\partial^2 \log L}{\partial \gamma_{1j}^2} &= -\alpha_{1j} \left( \frac{p_{1j} \log p_{1j} + (\log p_{1j})^2 p_{1j} - p_{1j}^2 \log p_{1j}}{(1 - p_{1j})^2} \right) + b_{1j} \log p_{1j} \\ &\quad + c_j \left( \frac{p_{1j}^2 \log p_{1j} + p_{1j} p_{2j} \log p_{1j} + p_{1j} p_{2j} (\log p_{1j})^2}{(p_{1j} + p_{2j})^2} \right), \end{aligned}$$

and for  $j = 1, \dots, k$ ,  $\frac{-\partial^2 \log L}{\partial \gamma_{1j} \partial \gamma_{1s}} = 0$ ,  $j \neq s$ .

$[-\partial^2 \Gamma_1]$  is a diagonal matrix. Similarly,  $[-\partial^2 \Gamma_2]$  is a diagonal matrix whose  $j$ th diagonal element is:

$$\begin{aligned} \frac{-\partial^2 \log L}{\partial \gamma_{2j}^2} &= -\alpha_{2j} \left( \frac{p_{2j} \log p_{2j} + (\log p_{2j})^2 p_{2j} - p_{2j}^2 \log p_{2j}}{(1 - p_{2j})^2} \right) + b_{2j} \log p_{2j} \\ &\quad + c_j \left( \frac{p_{2j}^2 \log p_{2j} + p_{1j} p_{2j} \log p_{2j} + p_{1j} p_{2j} (\log p_{2j})^2}{(p_{1j} + p_{2j})^2} \right), \end{aligned}$$

for  $j = 1, \dots, k$ . Additionally,

$$\begin{aligned} \frac{-\partial^2 \log L}{\partial \beta_r \partial \gamma_{1j}} &= \frac{p_{1j} \log p_{1j} (1 - p_{1j} + \log p_{1j})}{(1 - p_{1j})^2} \sum_{i \in D_j(d)} z_{ir} - \log p_{1j} \sum_{\ell \in C_j(d)} z_{\ell r} \\ &\quad + \frac{p_{1j} \log p_{1j} (p_{1j} + p_{2j} - p_{2j} \log p_{2j})}{(p_{1j} + p_{2j})^2} \sum_{\ell \in C_j} z_{\ell r}. \end{aligned}$$

$$\frac{-\partial^2 \log L}{\partial \beta_r \partial \gamma_{2j}} = \frac{p_{1j} p_{2j} \log p_{2j} \log p_{1j}}{(p_{1j} + p_{2j})^2} \sum_{\ell \in C_j(d)} z_{\ell r}$$

$$\frac{-\partial^2 \log L}{\partial \alpha_r \partial \gamma_{1j}} = \frac{-\partial^2 \log L}{\partial \beta_r \partial \gamma_{2j}}$$

$$\begin{aligned} \frac{-\partial^2 \log L}{\partial \alpha_r \partial \gamma_{1j}} &= \frac{p_{1j} \log p_{2j} (1 - p_{2j} + \log p_{2j})}{(1 - p_{2j}^2)} \sum_{i \in D_j(\bar{d})} z_{ir} \log p_{2j} \sum_{\ell \in C_j(\bar{d})} z_{\ell r} \\ &+ \frac{p_{2j} \log p_{2j} (p_{1j} + p_{2j} - p_{1j} \log p_{1j})}{(p_{1j} + p_{2j})^2} \sum_{\ell \in C_j} z_{\ell r} \end{aligned}$$

$$\begin{aligned} \frac{-\partial^2 \log L}{\partial \beta_r \partial \beta_q} &= \sum_{j=1}^k \left\{ \frac{p_{1j} \log p_{1j} (1 + \log p_{1j} - p_{1j})}{(1 - p_{1j})^2} \sum_{i \in D_j(d)} z_{ir} z_{iq} \right. \\ &\quad \left. - \log p_{1j} \sum_{\ell \in C_j(d)} z_{\ell q} z_{\ell r} - \frac{p_{1j} \log p_{1j} (p_{1j} + p_{2j} + p_{2j} \log p_{1j})}{(p_{1j} + p_{2j})^2} \sum_{\ell \in C_j} z_{\ell r} z_{\ell q} \right\} \end{aligned}$$

$$\begin{aligned} \frac{-\partial^2 \log L}{\partial \alpha_r \partial \alpha_q} &= \sum_{j=1}^k \left\{ \frac{p_{2j} \log p_{2j} (1 - p_{2j} + \log p_{2j})}{(1 - p_{2j}^2)} \sum_{i \in D_j(\bar{d})} z_{ir} z_{iq} \right. \\ &\quad \left. - \log p_{2j} \sum_{\ell \in C_j(d)} z_{\ell r} z_{\ell q} + \frac{p_{2j} \log p_{2j} (p_{1j} + p_{2j} + p_{1j} \log p_{2j})}{(p_{1j} + p_{2j})^2} \sum_{\ell \in C_j} z_{\ell r} z_{\ell q} \right\} \end{aligned}$$

$$\frac{-\partial^2 \log L}{\partial \alpha_r \partial \beta_q} = \sum_{j=1}^k \left\{ \frac{p_{1j} p_{2j} \log p_{1j} \log p_{2j}}{(p_{1j} + p_{2j})^2} \sum_{i \in C_j} z_{ir} z_{iq} \right\}$$

## 8. DISCUSSION

There remains much work to be done. These three models must be compared. The most useful model seems to be the semiparametric partial likelihood model, because software already exists for the estimation of its parameters. However, the simplified extension of Louis et al. settles down somewhat as the number of screenings increases. This suggests that, if sample sizes are large enough for one screen, they are also large enough for multiple screens. Of course, a power study would need to be performed to determine when sample sizes are large enough.

## 9. REFERENCES

1. Albert, A, Gertman, PM and Louis, TA (1978). Screening for the early detection of cancer I. The temporal natural history of a progressive disease state. Mathematical Biosciences 40, 1-59.
2. Albert, A, Gertman, PM, Louis, TA and Liu, S (1978). Screening for the early detection of cancer II. The impact of screening on the natural history of the disease. Mathematical Biosciences 40, 61-109.
3. Breslow, NE, Lubin, JH, Marek, P and Langholz, B (1983). Multiplicative models and cohort analysis. Journal of the American Statistical Association 78, 1-12.
4. Dinse, GE (1982). Nonparametric Estimation for Partially-Complete Time and Type of Failure Data. Biometrics 38, 417-431.
5. Dinse, GE (1988). Simple Parametric Analysis of Animal Tumorigenicity Data. Journal of The American Statistical Association 83, 638-649.
6. Dinse, GE and Lagakos, S (1982). Nonparametric Estimation of Lifetime and Disease Onset Distributions from Incomplete Observations. Biometrics 38, 921-932.
7. Louis, TA, Arthur, A and Heghinian, S (1978). Screening for the early detection of cancer III. Estimation of disease natural history. Mathematical Biosciences 40, 111-144.
8. Kalbfleisch, JD and Prentice, RL (1980). The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York.
9. Lawless, JF (1982). Statistical Models and Methods for Lifetime Data. Wiley, New York.
10. Mihalko, Daniel (1989). Final Report (April 1989), Contract #F33615-88-D-0609, Subcontract #SCEE-ARB/88-0002.
11. Rao, CR (1973). Linear Statistical Inference and Its Applications, Second Edition. John Wiley, New York.